UNIWORLD: A UNIFIED WORLD MODEL FOR REAL-TIME VISUAL WORLD SIMULATION

Anonymous authors
Paper under double-blind review



Figure 1: **UniWorld: from world geometry to continuous visual observations.** Conceptual overview: depth trajectories describe geometric evolution, while reference RGB specifies appearance for a shared video renderer.

ABSTRACT

Continuous visual world simulation requires coupling evolving scene geometry with temporally persistent appearance. We present UniWorld, a unified framework that integrates depth trajectories and reference appearance within a shared video generation model. The central design maintains visual and geometric histories on a common timeline. Noise-aligned overlapping continuation places generated history at the prescribed sampling noise level and commits only newly generated latents. Rolling depth latent state preserves the corresponding geometric controls, incrementally reads newly available geometry, and keeps the conditioning fixed within each sampling window. Together, these mechanisms ensure exact preservation of latent history while bounding the active depth state by the window size. We evaluate continuous generation, paired history interventions, and long-horizon behavior across diverse scene categories. The same generator supports estimated, native, and on-demand simulator depth, while cached and incremental execution paths produce bitwise-identical outputs. Appearance changes and geometric control branches operate through the same interface without modifying committed history. The measured steady-state DiT execution path achieves real-time generation throughput, demonstrating the efficiency of UniWorld for continuous visual world simulation.

1 INTRODUCTION

World models connect observations with environment dynamics, supporting simulation, prediction, and decision-making (Ha and Schmidhuber, 2018; Bruce et al., 2024; Zhu et al., 2026). Graphics engines and physical simulators provide structured and controllable scene states, whereas video generators provide rich visual appearance. Bridging these capabilities requires a renderer that can continuously transform evolving geometry into temporally consistent visual observations. We focus

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

on this simulator-to-renderer component: an upstream world source specifies a geometric trajectory, and a generative model produces the corresponding visual realization.

Different world sources expose distinct state variables, camera models, and control interfaces, which makes direct coupling to a shared visual generator difficult. We observe that the visible geometry produced by these sources can instead be represented in a common image-space form: depth. Image-aligned depth captures scene layout, occlusion, and geometric changes, while a reference RGB image specifies texture, color, and appearance semantics. This separation maps heterogeneous world states to a common rendering interface, enabling the same generator to operate on recorded scenes, simulator trajectories, and on-demand geometric inputs.

Continuous rendering introduces a second challenge: generation must remain consistent across finite sampling windows. Long-video methods extend temporal context, use overlapping windows, or compress history (Henschel et al., 2024; Zhang et al., 2025; Gao et al., 2025a), while autoregressive approaches condition future generation on earlier observations (Huang et al., 2025; Zhu et al., 2026). Depth-driven rendering must additionally preserve the geometric conditions associated with those observations. Continuous generation therefore requires two synchronized histories: the generated visual state and the depth state that produced it. Previously generated observations must remain fixed, the associated depth values must remain unchanged, and only newly requested geometry should produce newly committed outputs. A shared temporal contract makes these state transitions explicit and testable.

We propose UniWorld, a unified generative rendering framework that maintains visual and geometric states on a common timeline. *Noise-aligned overlapping continuation* places previously generated latents in the prefix of the next window and constrains them to the corresponding flow time throughout sampling. *Rolling depth latent state* preserves the matching geometric history, incrementally reads only newly available depth positions, and keeps the conditioning fixed within each window. After sampling, the system commits only the newly generated video suffix. The sampling window can therefore advance without rewriting committed latents or altering the controls associated with the overlapping region.

This design bounds the active state and per-window computation independently of the total sequence duration, thereby enabling continuous long-horizon rendering. We verify the state contract through actual model executions and evaluate UniWorld across four scene categories using estimated and native depth, controlled fixed-window interventions, and extended trajectories. An on-demand engine further demonstrates reference-appearance changes and branch-conditioned continuation through the same interface. The measured steady-state DiT execution path achieves real-time rendering throughput while maintaining a bounded active state during continuous generation.

Our contributions are:

1. **UniWorld.** A unified visual simulation framework that represents heterogeneous geometric trajectories using image-aligned depth and combines them with reference appearance, providing a common generative rendering interface for recorded scenes, simulators, and on-demand world trajectories.
2. **Noise-aligned overlapping continuation.** A sampling rule that constrains a generated multi-frame prefix at every flow time and commits only the newly generated suffix, thereby preserving previously generated latents while retaining temporal context for continued generation.
3. **Rolling depth latent state.** A synchronized geometric state that incrementally reads only newly available depth positions, preserves the conditions associated with committed observations, and maintains an active-state bound that is independent of sequence duration.

2 Related Work

2.1 Visual World Models

World Models learn compact representations of environment dynamics (Ha and Schmidhuber, 2018), and Genie learns interactive visual environments from video (Bruce et al., 2024). Astra combines autoregressive denoising, noisy histories, and action adapters to support exploration, manipulation, and driving (Zhu et al., 2026). These approaches link observations to actions or latent controls.

UniWorld complements these approaches by providing a shared visual rendering interface: an upstream simulator supplies the evolving geometry, and the generator continuously renders the corresponding appearance.

2.2 Conditional Video Generation

Large video models support conditioning on text and images (Wan Team, 2025; MiniMax, 2026). ControlNet, SparseCtrl, and CameraCtrl introduce structural or camera constraints (Zhang et al., 2023; Guo et al., 2023; He et al., 2024). LongVie and LongVie 2 combine multiple control signals for long-video synthesis (Gao et al., 2025a;b). Our implementation builds on H3, Video DeltaNet (VDN), Qwen3-VL, and the VideoX-Fun Union control branch (MiniMax, 2026; Xi et al., 2026; Bai et al., 2025; aigc-apps, 2026). The original image provides persistent semantic context, while a generated boundary image and temporally aligned depth serve as local conditioning inputs. Our modifications focus on how these conditioning inputs and the generated history are retained across windows.

2.3 Long-Video Generation

StreamingT2V combines short- and long-term appearance memory (Henschel et al., 2024), and FramePack compresses frame context (Zhang et al., 2025). Diffusion Forcing uses per-token noise levels during training, and Self Forcing addresses the mismatch between training histories and model-generated histories (Chen et al., 2024; Huang et al., 2025). RePaint conditions diffusion sampling on known regions by replacing noisy observations (Lugmayr et al., 2022). LongVie uses unified noise initialization and global control normalization (Gao et al., 2025a). Building on these principles, we focus on a joint temporal contract for a flow renderer that uses eight evaluations. Under this contract, the multi-frame target prefix is matched to the current sampling noise level, canonical latent values are preserved in overlapping depth segments, and incremental output is aligned with the global VAE phase.

3 Approach

3.1 Task Definition and Preliminaries

The method takes as input a single reference image I_{ref}, optional text p, and a spatially and temporally aligned depth sequence D, and outputs a video $\hat{V}$ at playback rate f. The target generation rate for real-time operation is 14 FPS. For subsequent windows, boundary images are obtained from previously generated content. All temporal intervals use zero-based, half-open indexing. Let N and F denote the aligned temporal lengths in pixel space and latent space, respectively. Window i spans $\Omega_i = [a_i, b_i)$ in latent time, with length $L_i = b_i - a_i$, inherited history H_i of length h_i, depth condition C_i, boundary image I_i, and a slice ϵ_i of the global noise. The first window has $h_0 = 0$.

$$c_{\mathrm{g}} = E_{\mathrm{Qwen}}(I_{\mathrm{ref}}, p), \qquad \hat{Z}_i = G_\theta^{(K)}(H_i, C_i, I_i, c_{\mathrm{g}}, \epsilon_i), \qquad K = 8. \tag{1}$$

Here $G_\theta^{(K)}$ includes the projection and sampler described below. The windows are generated sequentially because each window depends on previously generated history, whereas generation within each window remains bidirectional.

3.2 UniWorld: A Unified Framework for Visual World Simulation

For source m, let $\xi^{(m)}$ denote its state or observations, and let $\kappa^{(m)}$ denote the corresponding camera and alignment information. A source adapter $\mathcal{A}_m$ produces visible depth, while $\mathcal{N}_m$ standardizes its representation:

$$\bar{D}^{(m)} = \mathcal{N}_m\Big(\mathcal{A}_m(\xi^{(m)}, \kappa^{(m)})\Big), \qquad C = \mathrm{Patchify}\big[\mathrm{Norm}_z\big(E_{\mathrm{VAE}}(\mathrm{RGBReplicate}(\bar{D}))\big)\big]. \tag{2}$$

Time, viewpoint, crop, near/far direction, scale, and encoding phase remain consistent throughout the trajectory. We replicate grayscale depth across the RGB channels, use the VAE posterior mode with FP32 latent normalization, and patchify with shape $(1, 2, 2)$. Each latent time position contains $S = (H_{\mathrm{px}}/32)(W_{\mathrm{px}}/32)$ spatial tokens, each with dimension 96. Both estimated and native depth

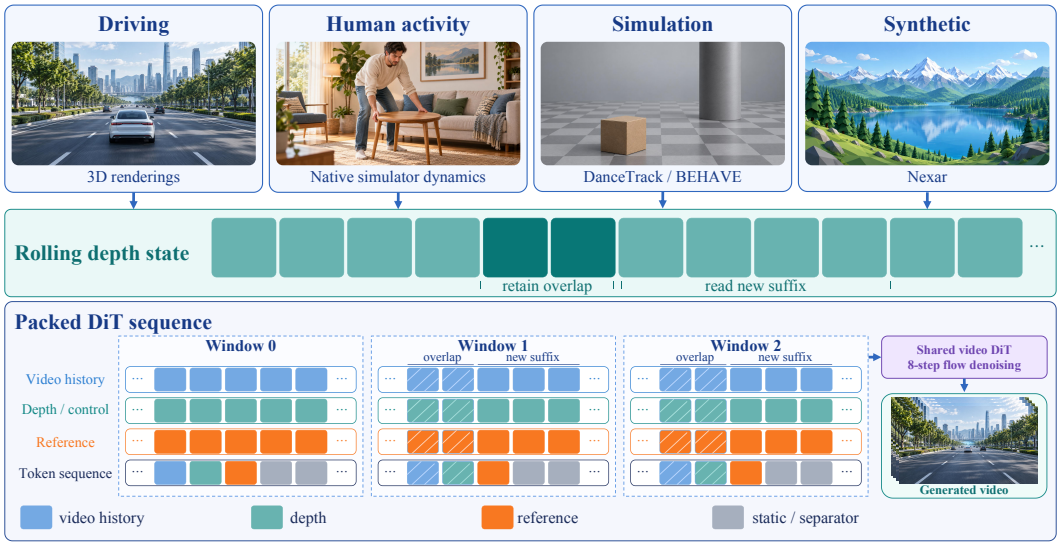

Figure 2: **UniWorld: a unified renderer for heterogeneous world sources.** Driving, human-activity, simulation, and synthetic sources are converted into a common rolling depth state. Each continuation window packs video history, depth/control, reference, and separator tokens into a shared DiT sequence. Overlapping positions are retained, only the newly required depth suffix is read, and only the newly generated video suffix is committed to the output stream.

follow the verified far-bright convention of the checkpoint. An online adapter encodes newly requested geometry under the same global phase; Appendix C.2 specifies the source mappings.

Appearance conditioning is provided through two paths (Figure 2). The original image and text are encoded once by Qwen3-VL and reused throughout the trajectory. A separate VAE first-frame condition uses $I_0 = I_{\text{ref}}$ initially and generated RGB at the start of each subsequent window. These independent conditioning rows are distinct from the multi-frame history, which occupies the target video sequence itself.

3.3 Noise-Aligned Overlapping Continuation

History as a sampling boundary condition. We adopt the noise-to-data flow convention (Lipman et al., 2023):

$$z_t = tz_{\text{clean}} + (1-t)\epsilon, \qquad v = z_{\text{clean}} - \epsilon, \qquad 0 \le t \le 1. \tag{3}$$

For window i, let M_i mask the first h_i latent positions corresponding to inherited history, and extend H_i by zeros outside this prefix. We impose the history through the projection

$$\boxed{\mathcal{P}_{i,t}(z) = M_i \odot \left[tH_i + (1-t)\epsilon_i\right] + (1-M_i) \odot z.} \tag{4}$$

Thus, the inherited prefix follows the same noise-to-data trajectory as the current sample instead of being inserted as a clean latent into a noisy state. Global positions shared by overlapping windows use the same realization of global noise.

On the shifted flow grid $0 = t_0 < \cdots < t_K = 1$, we initialize

$$z_i^{(0)} = \mathcal{P}_{i,t_0}(\epsilon_i), \tag{5}$$

and perform

$$\tilde{z}_i^{(k+1)} = z_i^{(k)} + (t_{k+1} - t_k) v_\theta\left(z_i^{(k)}, t_k; \mathcal{C}_i\right), \qquad z_i^{(k+1)} = \mathcal{P}_{i,t_{k+1}}\left(\tilde{z}_i^{(k+1)}\right), \tag{6}$$

where

$$\mathcal{C}_i = (C_i, I_i, c_{\text{g}}) \tag{7}$$

collects the fixed depth, boundary-image, and global appearance conditions for window i.

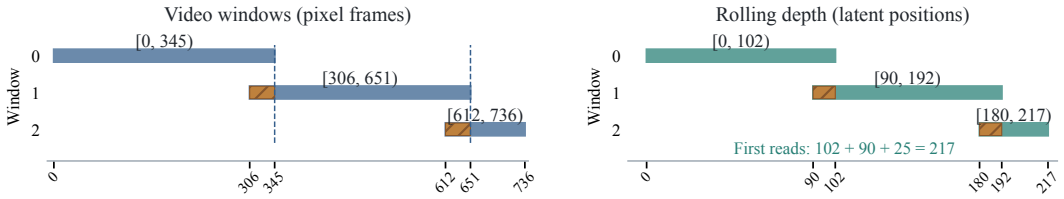

Figure 3: **Aligned visual and geometric state.** Hatched intervals denote inherited history. Depth is read exactly once over $[0, 102)$, $[102, 192)$, and $[192, 217)$. Dashed pixel lines mark the nominal latent-commitment boundaries, while stable incremental pixel output retains an additional five-frame blend tail.

The projection ensures that the inherited prefix is aligned with the current flow noise level before every model evaluation. After each numerical update, the predicted prefix is replaced by its prescribed history trajectory, whereas the predicted suffix is retained. Consequently, the inherited history fully participates in the bidirectional forward pass within each window and guides newly generated content without drifting. At $t_K = 1$, the projected prefix exactly recovers H_i. Because the projection is an element-wise replacement, it requires no additional model evaluations.

History preservation and output commitment. For fixed history, mask, and noise, the projection is idempotent. The initialization and equation 6 imply

$$\mathcal{P}_{i,t}(\mathcal{P}_{i,t}(z)) = \mathcal{P}_{i,t}(z), \qquad M_i \odot z_i^{(k)} = M_i \odot [t_k H_i + (1 - t_k)\epsilon_i]. \tag{8}$$

Thus, the terminal prefix equals H_i. We preserve the previously committed sequence $\hat{Z}_{<i}$ and append only the newly generated suffix:

$$\hat{Z}_{\leq i} = \mathrm{concat}\big(\hat{Z}_{<i}, z_i^{(K)}[h_i : L_i]\big). \tag{9}$$

Appendix A proves exact latent-history preservation. Section 4.4 evaluates the generated suffix under fixed-window interventions, while the decoder described below preserves the global phase during RGB commitment.

Window geometry and boundary images. Under the temporal geometry of the video VAE, $N = 17n+5$ pixel frames correspond to $F = 5n+2$ latent frames. The default pixel window, overlap, and stride are $(W, O, \Delta) = (345, 39, 306)$, corresponding to $(L_{\max}, h, \Delta_{\text{latent}}) = (102, 12, 90)$. For the evaluation stream at 24 FPS, a 30-second request is internally padded to 736 pixel frames and 217 latents, and the decoded output is trimmed to 720 frames. The corresponding latent windows are $[0, 102)$, $[90, 192)$, and $[180, 217)$ (Figure 3). The subsequent windows start at pixel frames 306 and 612, while the nominal new-suffix boundaries are 345 and 651.

The next boundary RGB is taken at the *start of the next window*: decoding latents $[85, 97)$ and selecting local pixel frame 17 recovers frame 306, matching full-window decoding. RGB export preserves a single global VAE phase. Incremental delivery retains five unfinished blend frames and emits $[0, 340)$, $[340, 646)$, and then $[646, 720)$; these pixel-delivery boundaries differ from the latent commitment boundaries.

3.4 Rolling Depth Latent State

For $i > 0$, let $h_i = b_{i-1} - a_i$ denote the overlap with the preceding window. Rather than reprocessing overlapping geometry, we retain the corresponding canonical depth latents and append only newly requested positions:

$$\boxed{C_i = \mathrm{concat}\left(C_{i-1}[a_i - a_{i-1} : L_{i-1}],\ \mathrm{ReadDepth}(b_{i-1}, b_i)\right).} \tag{10}$$

The first window reads $[a_0, b_0)$. A state update is committed only after interval, shape, dtype, and finite-value checks pass; retrying a window that has already been committed returns the cached condition. Once constructed, the depth state remains fixed throughout all K model evaluations:

$$C_i^{(0)} = C_i^{(1)} = \cdots = C_i^{(K-1)} = C_i. \tag{11}$$

Algorithm 1 UniWorld continuous generation

Input: I_{ref}, p, depth source $\mathcal{D}$, windows $\{\Omega_i\}$, global noise ϵ, grid $\{t_k\}_{k=0}^{K}$
$c_{\mathrm{g}} \leftarrow E_{\mathrm{Qwen}}(I_{\mathrm{ref}}, p)$; $I_0 \leftarrow I_{\mathrm{ref}}$; $Z \leftarrow [\,]$; depth state $\leftarrow [\,]$
for each $\Omega_i = [a_i, b_i)$ **do**
  $C_i \leftarrow \mathrm{RollDepth}(\mathcal{D}, \Omega_i)$
  $H_i \leftarrow Z[a_i : a_i + h_i]$; $\epsilon_i \leftarrow \epsilon[a_i : b_i]$
  $\mathcal{C}_i \leftarrow \big(C_i, \mathrm{EncodeBoundary}(I_i), c_{\mathrm{g}}\big)$
  $z_i \leftarrow \mathcal{P}_{i,t_0}(\epsilon_i)$
  for $k = 0, \ldots, K-1$ **do**
    $v_i \leftarrow v_\theta(z_i, t_k; \mathcal{C}_i)$
    $z_i \leftarrow \mathcal{P}_{i,t_{k+1}}(z_i + (t_{k+1} - t_k)v_i)$
  end for
  $Z \leftarrow \mathrm{concat}\big(Z, z_i[h_i : L_i]\big)$; depth state $\leftarrow C_i$
  if another window remains: $I_{i+1} \leftarrow \mathrm{DecodeAtNextStart}(Z)$
end for
Decode Z under the global VAE phase and trim alignment padding to the requested delivery length.

Thus, unlike video history, depth does not follow the flow noise level; it remains a clean FP32 structural condition. Nevertheless, features that jointly depend on depth, the noisy video state, and the current flow time are recomputed at every model evaluation.

Under sequential, gap-free traversal and correct source reads, three properties follow (Appendix A). First,

$$C_i = C[a_i : b_i], \tag{12}$$

so overlapping depth positions retain exactly the same canonical latent values and dtype across windows. Second, strictly increasing right endpoints produce disjoint first-read intervals with a total temporal extent of F, independent of the number of flow evaluations K. Therefore, each global depth position is introduced into the rolling state exactly once. Third, the active depth rows satisfy

$$\boxed{M_{\mathrm{depth}} \le 4L_{\max}S \times 96 \text{ bytes.}} \tag{13}$$

For the default $L_{\max} = 102$, this yields an upper bound of $19.050\,\mathrm{MiB}$ at 960×544 resolution and $9.413\,\mathrm{MiB}$ at 672×384. The bound covers only the active canonical depth rows and excludes transient copies, fused conditioning features, model activations, and storage owned by the underlying geometry source. Relative to direct slicing of the same canonically encoded source, rolling depth preserves the conditioning tensor exactly. Its role is therefore to organize the active state and avoid redundant access rather than to alter the model output.

3.5 Joint Generation Algorithm

The logical state of window i is

$$\mathcal{S}_i = (\Omega_i, H_i, C_i, I_i, \epsilon_i, c_{\mathrm{g}})\,. \tag{14}$$

A single global temporal plan determines the indices of video history, depth state, and noise, ensuring that all three refer to the same latent-time positions. Algorithm 1 combines rolling depth, noise-aligned history preservation, and boundary-image updates into a unified continuous generation procedure.

4 Experiments

4.1 Experimental Setup

Data and implementation. The primary set contains 100 thirty-second clips, with 25 each from driving, human activity, simulation, and synthetic scenes. Video Depth Anything (VDA) ViT-L FP32 estimates (Chen et al., 2025) are prepared at 832×480, 16 fps, normalized over each trajectory, and encoded for 960×544 generation. GT follows source timestamps at 24 fps and the same canvas

Table 1: VBench quality ($\times 100$, $\uparrow$); best in bold.

Configuration	Quality Score	Subject cons.	Background cons.	Temporal flickering	Motion smoothness	Dynamic degree	Aesthetic quality	Imaging quality
H3 50-step	79.87	**87.46**	92.94	**96.72**	98.47	83.50	52.60	62.28
LongVie2	78.54	86.90	92.86	96.46	97.83	85.00	50.50	58.64
LongVie AR2	78.62	86.28	92.19	96.63	97.84	**86.00**	50.06	60.24
H3/VDN DMD8	**80.54**	87.20	**93.10**	96.54	**98.68**	85.50	**54.57**	**63.51**

Table 2: Thirty-second fidelity and depth consistency, 960×544, $W/O = 345/39$, 8 NFE. Rank and Edge denote depth Spearman and edge F1; NMAE fixes calibration after five seconds. The final row excludes configured-manifest overlap.

Set	n	PSNR↑	SSIM↑	LPIPS↓	Rank↑	Edge↑	NMAE↓
Driving	25	13.4429	0.5109	0.4727	0.9042	0.7027	0.0554
Human activity	25	14.9858	0.4799	0.3883	0.9299	0.7509	0.0523
Simulation	25	12.4034	0.3757	0.5914	0.7665	0.5223	0.2524
Synthetic	25	12.7571	0.3612	0.5475	0.9117	0.5655	0.0877
All	100	13.3973	0.4319	0.5000	0.8781	0.6354	0.1120
No manifest overlap	90	13.1587	0.4357	0.5127	0.8702	0.6212	0.1181

transform for fidelity evaluation; real-time throughput is evaluated against a 14-FPS target. Additional studies cover short-video quality, 24 sources under nine paired conditions, 25 native-depth inputs, 30 long videos, and three PyBullet conditions. Appendix C records source identities, splits, and configured-manifest sensitivity analyses.

The generator combines the released MiniMax-H3 ControlNet-Union for native depth conditioning with the VDN-H3 8-step model obtained by DMD distillation. The Union branch connects to five of 50 H3 blocks; inference uses eight evaluations per window, shifts 12/3, and guidance 1. A 30-second output spans three windows and 720 RGB frames. Each worker uses eight H100 GPUs with FSDP/Ulysses, with no additional joint training of the released components.

Metrics and statistics. We report video-level RGB PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018). Geometry uses VDA rank correlation, depth-edge F1, and NMAE with calibration fixed after five seconds; native comparisons share inverse-metre GT and original valid masks. Temporal error measures disagreement with GT frame changes at fixed seams. Intervals use 10,000 category-stratified source-group bootstrap replicates, one fixed seed per source, and marginal 95% intervals. Appendix C specifies sampling, calibration, and aggregation.

4.2 Video Quality Comparison

We compare the H3/VDN DMD8 generator used by UniWorld with H3 50-step, LongVie2, and LongVie AR2 on the same 100 sources. Table 1 averages 5- and 14-second outputs equally (200 videos per configuration), using VBench's seven quality dimensions and officially weighted Quality Score (Huang et al., 2024). DMD8 achieves the highest Quality Score, background consistency, motion smoothness, aesthetic quality, and imaging quality, exceeding H3 50-step by 0.67 Quality Score points. Appendix C.8 gives native configurations, duration-level results, and paired intervals.

4.3 Multi-Scenario Continuous Generation

One checkpoint completes all 100 clips and 300 windows. Table 2 reports source fidelity and depth consistency across the four categories; overall PSNR is 13.3973 dB (95% CI $[12.9211, 13.8794]$). Figure 4 shows outputs from the first frozen ablation identity in each category. Appendix C provides full intervals and continuation-only measurements.

Native depth. The same model also completes all 25 PointOdyssey/TartanAir inputs with native metric depth by changing only the source adapter. Appendix C.2 and Table S3 give the matched comparison against common native GT, paired intervals, and valid-support rules.



Figure 4: **One model across four scene categories.** Actual reference, GT, and generated frames at the first continuation boundary and final frame. Rows use the first frozen ablation identity per category; selection is independent of quality scores.

Table 3: Paired effects, 24 sources per arm; brackets are marginal 95% bootstrap intervals. PSNR covers $[345, 720)$; LPIPS and temporal error cover fixed seams. Positive LPIPS/error differences favor the full method.

Arm minus full	ΔPSNR (dB)↑	ΔLPIPS↓	ΔE_{temp} ↓
Boundary RGB only	$0.00637\ [-0.01467, 0.02591]$	$0.00195\ [0.00033, 0.00362]$	$-0.00047\ [-0.00140, 0.00039]$
Endpoint only	$0.00637\ [-0.01467, 0.02591]$	$0.00195\ [0.00033, 0.00362]$	$-0.00047\ [-0.00140, 0.00039]$
Fresh history noise	$-0.01337\ [-0.08561, 0.05934]$	$-0.00198\ [-0.00344, -0.00045]$	$0.00021\ [-0.00054, 0.00103]$
Last 6 history latents	$0.00623\ [-0.00484, 0.01751]$	$0.00030\ [-0.00010, 0.00072]$	$0.00002\ [-0.00022, 0.00029]$
Last 2 history latents	$0.00917\ [-0.00888, 0.02600]$	$0.00113\ [-0.00016, 0.00261]$	$-0.00023\ [-0.00056, 0.00008]$
Direct global slice	$0.00000\ [0.00000, 0.00000]$	$0.00000\ [0.00000, 0.00000]$	$0.00000\ [0.00000, 0.00000]$
Window normalization	$-0.00897\ [-0.04902, 0.03094]$	$0.00063\ [-0.00076, 0.00200]$	$0.00085\ [-0.00014, 0.00210]$
Aligned re-encoding	$0.00000\ [0.00000, 0.00000]$	$0.00000\ [0.00000, 0.00000]$	$0.00000\ [0.00000, 0.00000]$

4.4 Mechanism Analysis and Ablations

State consistency in real model runs. Full-method replay, rolling/direct reads, and aligned depth re-encoding match exactly in all 24 cases; all three online/cached pairs also match. Appendix C.7 and Table S9 retain the complete state checks, including committed history under a control branch.

Fixed-window comparisons. We select six distinct source groups per category by a fixed deterministic ordering. All nine conditions share the exact first-window output, suffix noise, reference/text, window geometry, commitment boundaries, schedule, and VAE. Partial-history variants constrain the final 6 or 2 of the 12 overlap latents, leaving the discarded overlap unchanged. Table 3 reports each alternative minus full per-step projection. Full projection reduces seam LPIPS by 0.00195 relative to boundary RGB only in this paired sample. Continuation PSNR and seam temporal-error intervals include zero, and fresh history noise gives slightly lower seam LPIPS. The endpoint-only control matches boundary-only output exactly: both have unconstrained history during model evaluations and discard the predicted overlap. This isolates per-step history participation from final output preservation. Appendix C.6 provides continuation, geometry, and manifest-exclusion results under the same protocol.

4.5 Runtime and Resource Usage

DiT forward throughput. Four fixed sources at each resolution receive one full-window warmup and three hot repeats: 32 requests including warmups, 12 hot requests per canvas. We time actual H3/VDN and Union forwards, including FSDP/Ulysses communication, with CUDA synchronization and per-step maxima over eight ranks. The devices are isolated from other experiment jobs; final latents match the corresponding complete-generation runs.

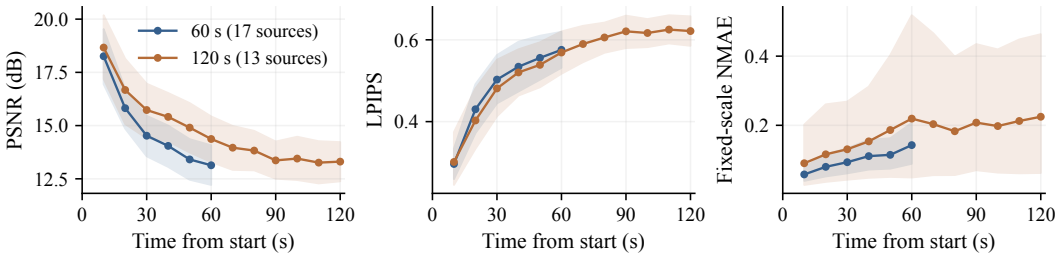

Figure 5: **Continuous generation over 60–120 seconds.** Ten-second aggregates with category-stratified source-group 95% bootstrap bands. The duration groups use different sources; the 120-second group is a training-split stress test.

We evaluate real-time throughput against a 14-FPS target. Full-continuation windows produce 14.472 and 21.361 new FPS at 960×544 and 672×384, respectively, meeting the target at both resolutions. These timings cover synchronized H3/VDN and Union forwards with FSDP/Ulysses communication on eight H100 GPUs; Appendix D reports the complete window profiles and whole-trajectory throughput.

Depth-state resources. Active depth rows occupy 19.050 MiB at 960×544 and 9.413 MiB at 672×384, independently of trajectory length. Rolling reads 217/427/852 positions for 30/60/120-second outputs, versus 241/475/960 for direct slicing, while preserving identical output. This provides a bounded geometric working state and avoids repeated source requests. Consumer copies, carrier, hints, model storage, and growing global caches are accounted for separately, together with measured reader-switch costs, in Appendix D.2.

4.6 Long-Horizon Behaviour and Applications

The same window interface advances through 17 sixty-second videos from existing validation/test splits and thirteen 120-second training-split stress tests. Their full-video PSNR/LPIPS are 14.4542/0.49032 and 14.4277/0.54645. Figure 5 characterizes fidelity and fixed-calibration geometric agreement over successive ten-second blocks; each duration uses its own source set and source-group confidence bands.

Three procedural PyBullet conditions (Coumans and Bai, 2016–2021) exercise warm/cool reference appearance, a camera revisit at frame 480, and a control branch activated at the second-window request, affecting frame 357 onward. On-demand and cached depth give bitwise-identical final latents in all three pairs. The control branch preserves the first window exactly and changes the new suffix (latent mean absolute difference 0.039076). Warm/cool references share identical geometry while producing different appearance. The first request reads 357 geometry frames, and RGB is delivered incrementally with the canonical decoder phase. Appendix D.4 reports the appearance, revisit, and response measurements for this window-level engine interface.

5 Conclusion

We presented UniWorld, a unified framework for continuous visual world simulation from depth trajectories and reference appearance. Its joint state contract preserves committed video latents, retains matching geometric controls, and reads new depth positions incrementally with bounded active state. One shared generator connects estimated depth, native simulation data, and an on-demand engine; exact replay and cached/online equivalence verify this connection in real model runs. Multi-scenario generation, fixed-window interventions, and long-horizon trajectories characterize the resulting visual simulation, while stage-specific measurements make its computation and data-access costs explicit. The measured steady-state DiT path achieves real-time throughput while maintaining bounded active state for continuous generation.

REFERENCES

aigc-apps. VideoX-Fun: A video generation pipeline for diffusion transformer, 2026. URL https://github.com/aigc-apps/VideoX-Fun. Software and model release.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL technical report, 2025. URL https://arxiv.org/abs/2511.21631. arXiv preprint arXiv:2511.21631.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktaschel. Genie: Generative interactive environments. *arXiv preprint arXiv:2402.15391*, 2024. URL https://arxiv.org/abs/2402.15391.

Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *arXiv preprint arXiv:2407.01392*, 2024. URL https://arxiv.org/abs/2407.01392.

Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video Depth Anything: Consistent depth estimation for super-long videos. *arXiv preprint arXiv:2501.12375*, 2025. URL https://arxiv.org/abs/2501.12375.

Erwin Coumans and Yunfei Bai. PyBullet, a python module for physics simulation for games, robotics and machine learning, 2016–2021. URL https://pybullet.org. Experiments use PyBullet 3.2.7 TinyRenderer.

Jianxiong Gao, Zhaoxi Chen, Xian Liu, Jianfeng Feng, Chenyang Si, Yanwei Fu, Yu Qiao, and Ziwei Liu. LongVie: Multimodal-guided controllable ultra-long video generation. *arXiv preprint arXiv:2508.03694*, 2025a. URL https://arxiv.org/abs/2508.03694.

Jianxiong Gao, Zhaoxi Chen, Xian Liu, Junhao Zhuang, Chengming Xu, Jianfeng Feng, Yu Qiao, Yanwei Fu, Chenyang Si, and Ziwei Liu. LongVie 2: Multimodal controllable ultra-long video world model. *arXiv preprint arXiv:2512.13604*, 2025b. URL https://arxiv.org/abs/2512.13604.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. SparseCtrl: Adding sparse controls to text-to-video diffusion models. *arXiv preprint arXiv:2311.16933*, 2023. URL https://arxiv.org/abs/2311.16933.

David Ha and J"urgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. URL https://arxiv.org/abs/1803.10122.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. URL https://arxiv.org/abs/2404.02101.

Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. StreamingT2V: Consistent, dynamic, and extendable long video generation from text. *arXiv preprint arXiv:2403.14773*, 2024. URL https://arxiv.org/abs/2403.14773.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025. URL https://arxiv.org/abs/2506.08009.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2311.17982.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.02747.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using denoising diffusion probabilistic models. *arXiv preprint arXiv:2201.09865*, 2022. URL https://arxiv.org/abs/2201.09865.

MiniMax. MiniMax-H3, 2026. URL https://huggingface.co/MiniMaxAI/MiniMax-H3. Model release.

Wan Team. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314. arXiv preprint arXiv:2503.20314.

Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Haocheng Xi, Yiming Xie, Hexu Zhao, Yiwen Zhang, Michael Liu, Thomas Creavin, Kurt Keutzer, Xiuyu Li, Zhaoyang Lv, Chenfeng Xu, and Haiwen Feng. Video DeltaNet: A video-native hybrid attention for livestream video generation. *arXiv preprint arXiv:2609.20744*, 2026. URL https://arxiv.org/abs/2609.20744.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. URL https://arxiv.org/abs/2302.05543.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. *arXiv preprint arXiv:2504.12626*, 2025. URL https://arxiv.org/abs/2504.12626.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. URL https://arxiv.org/abs/1801.03924.

Yixuan Zhu, Jiaqi Feng, Wenzhao Zheng, Yuan Gao, Xin Tao, Pengfei Wan, Jie Zhou, and Jiwen Lu. Astra: General interactive world model with autoregressive denoising. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=8UZpmrxoLG.